Victor Meng ID: 52057282
Ryan Hahn ID: 26443671
Justine Woo ID: 60622683

Assignment 3: M3

PERFORMS POORLY

1. **the of for with**: slow retrieval
2. **computer science**: slow retrieval
3. **Emily Navarro**: low relevance
4. **hello world**: low relevance
5. **software engineering**: low content pages
6. **ics 46**: low relevance
7. **home**: low relevance
8. **class**: near duplicate results
9. **kill murder**: low relevance
10. **in4matx**: low relevance

PERFORMS WELL

11. taco bell
12. professor
13. information retrieval
14. music
15. Pattis
16. JavaScript
17. Python
18. data science
19. faculty
20. ics 45c

HOW WE IMPROVED THE ENGINE

- Low relevance
  - Fields: boost the score of postings with the word in an “important” tag
  - URL analysis: boost the score of documents that have query terms in the URL
  - Anchor text: boost the score of documents who have other websites linking to them with anchor text containing query terms
  - AND condition: boost the score of documents that have all query terms
- Low content pages
  - PageRank: boost the score of pages with many incoming and outgoing links
- Duplicate results
  - Simhash: ignore near duplicate documents during indexing
- Slow retrieval
  - Cosine Vector Scoring Model: faster than boolean AND
  - Retrieval Batching: select the top 1000 documents for each word in the query to present to the user first
  - Filtering stop words: ignore the stop words in queries if not necessary, otherwise reduce the number of postings retrieved for them to save computation time